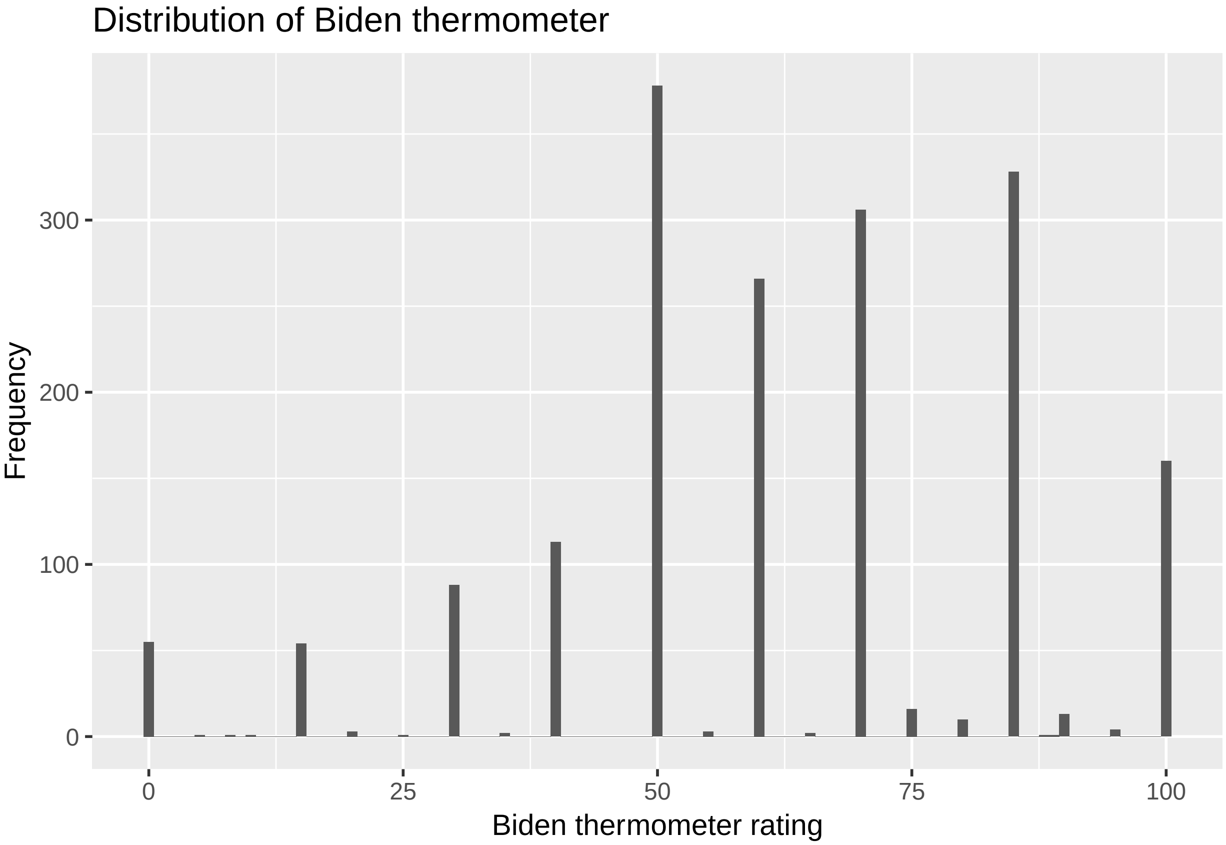
Homework 02: Sexy Joe Biden

Ellen Hsieh

1.

```
biden = read.csv('/Users/ellenhsieh/Documents/UChicago/2019 Winter/Modeling/hw02/data/nes2008.csv')

qplot(biden$biden,
      geom="histogram",
      binwidth = 1,
      main = "Distribution of Biden thermometer",
      xlab = "Biden thermometer rating",
      ylab = "Frequency"
      )
```



The distribution is skewed to the left. Although the respondents can choose any number to present their feeling thermometer from 0-100, most of them choose the number divisible by 5 or 10, not any random number like 37.

2.

```
model1 = lm(biden ~ age, data = biden)
summary(model1)
```

```
##
## Call:
## lm(formula = biden ~ age, data = biden)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.876 -12.318  -1.257  21.684  39.617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.19736    1.64792   35.92   <2e-16 ***
## age          0.06241    0.03267    1.91   0.0563 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.44 on 1805 degrees of freedom
## Multiple R-squared:  0.002018,   Adjusted R-squared:  0.001465
## F-statistic: 3.649 on 1 and 1805 DF,  p-value: 0.05626
```

```r
cor(biden$biden, biden$age)
```

```
## [1] 0.04491797
```

a. Yes, there is a relationship between age and biden feeling thermometer.

b. The relationship is pretty weak. When the age increases by one, biden feeling thermometer increases only by 0.062.

c. It's positive.

d. The R^2 for this model is 0.002, which indicates that this age model can only explain 0.002% of the variance in the biden feeling thermometer.

e.

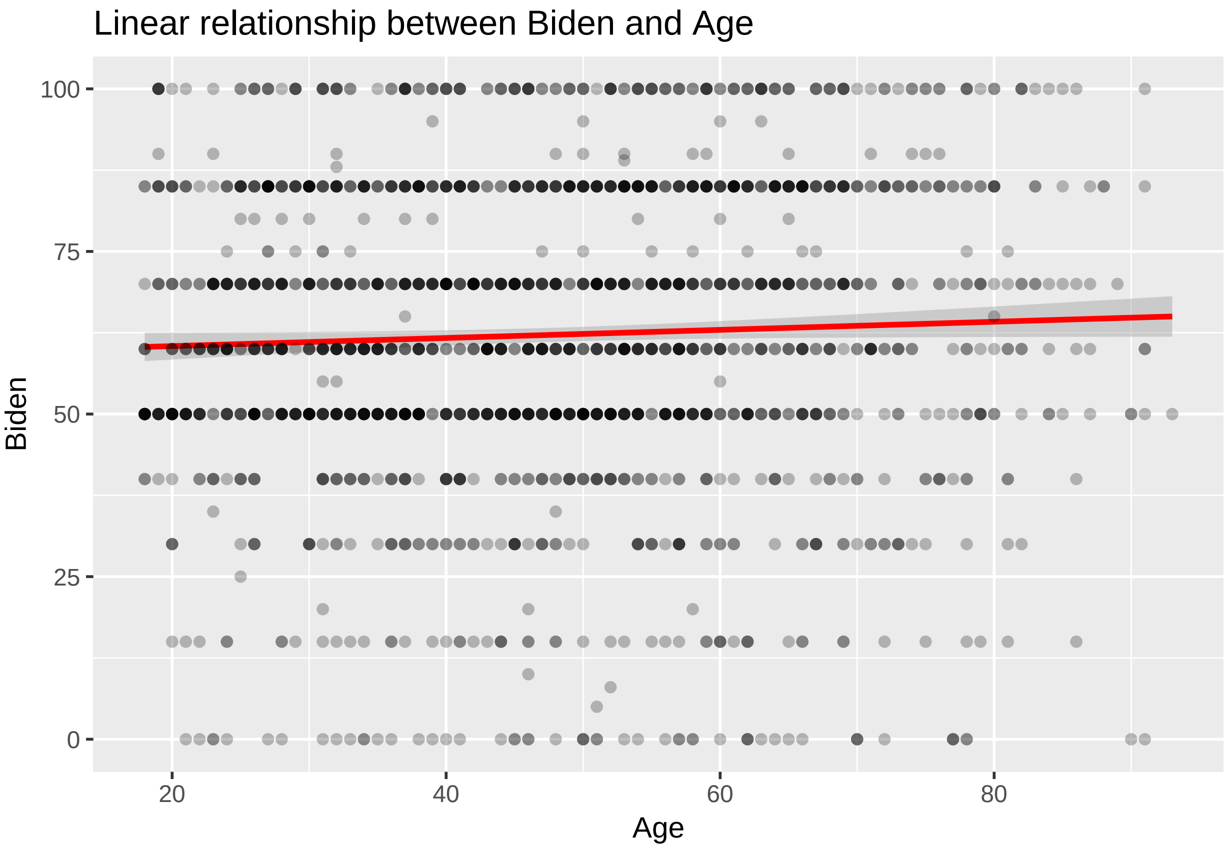
```r
predict(model1, data.frame(age=45), level = 0.95, interval = 'confidence')
```

```
##       fit      lwr      upr
## 1 62.0056 60.91177 63.09943
```

The 95% confidence interval for the predicted biden at age 45 is between 60.9 and 63.1.

f.

```r
ggplot(biden, aes(x = age, y = biden)) +
  stat_smooth(method = "lm", col = "red") +
  geom_point(alpha = .25) +
  scale_y_continuous(limits = c(0, 100)) +
  labs(title = "Linear relationship between Biden and Age",
        x = "Age",
        y = "Biden")
```



3.

```
model2 = lm(biden ~ age + female + educ, data = biden)
summary(model2)
```

```
##
## Call:
## lm(formula = biden ~ age + female + educ, data = biden)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.084 -14.662   0.703  18.847  45.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.62101    3.59600  19.083  < 2e-16 ***
## age          0.04188    0.03249   1.289    0.198
## female       6.19607    1.09670   5.650 1.86e-08 ***
## educ        -0.88871    0.22469  -3.955 7.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.16 on 1803 degrees of freedom
## Multiple R-squared:  0.02723,    Adjusted R-squared:  0.02561
## F-statistic: 16.82 on 3 and 1803 DF,  p-value: 8.876e-11
```

a. In this model, gender and education are strongly related to biden feeling thermometer. However, age is not statistically significant(p-value is 0.2 > 0.05)

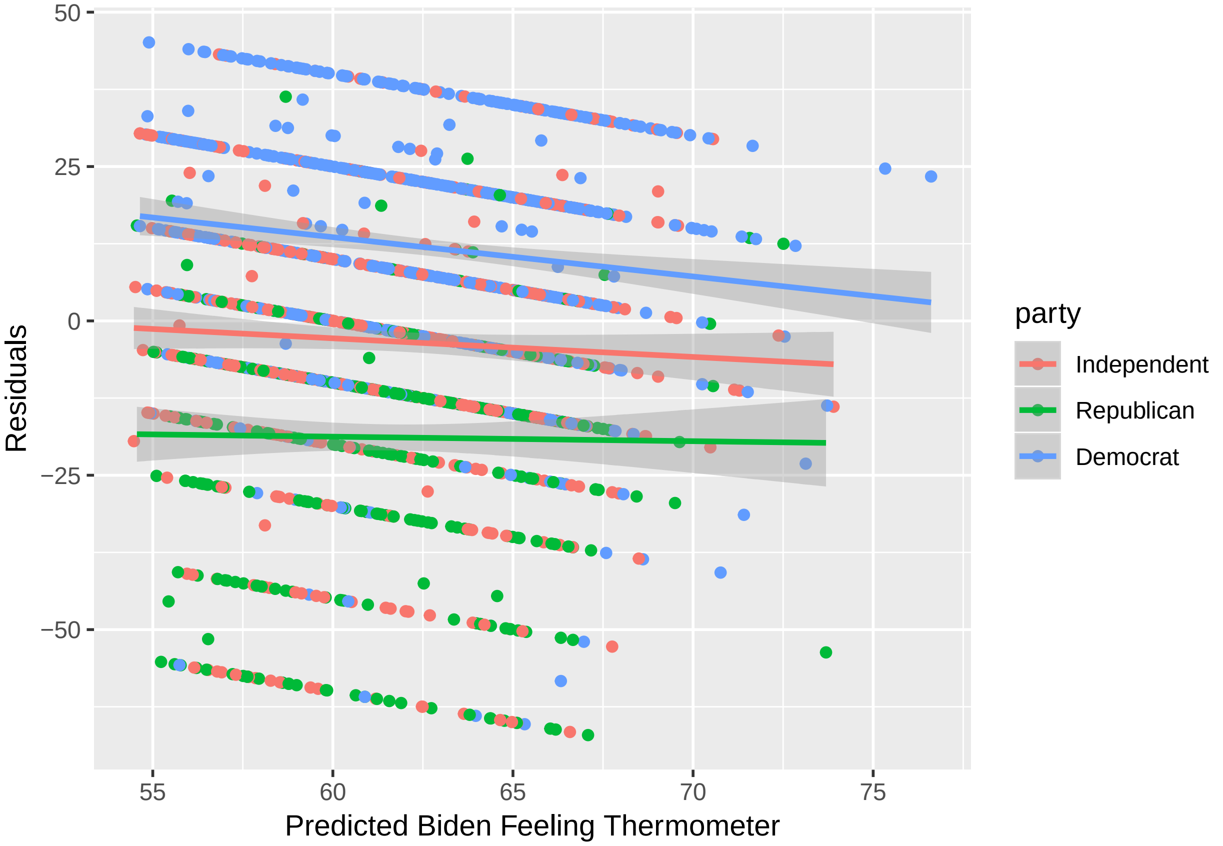
b. This gender parameter suggests that women have marginally higher feeling towards Biden than men.

c. This age-gender-education model explains 0.027% of the variance in biden feeling thermometer, which is higher than the previous one(0.002%). Thus, this model is better than the age-only model.

d.

```
biden$fitted2 = fitted(model2)
biden$residual2 = residuals(model2)
biden$party = interaction(factor(biden$rep), factor(biden$dem))
levels(biden$party) = c('Independent', 'Republican', 'Democrat', 'both')
biden$party = droplevels(biden$party, exclude = 'both')

ggplot(biden, aes(fitted2, residual2, color = party)) + geom_point(aes(color = party)) + stat_smooth(aes
labs(x = "Predicted Biden Feeling Thermometer",
     y = "Residuals")
```



The residuals for a good model should be scattered randomly, however, the residuals in this model are scattered in certain pattern. Also, the residuals for Democrat is higher than the other two.

4.

```
model3 = lm(biden ~ age + female + educ + dem + rep, data = biden)
summary(model3)
```

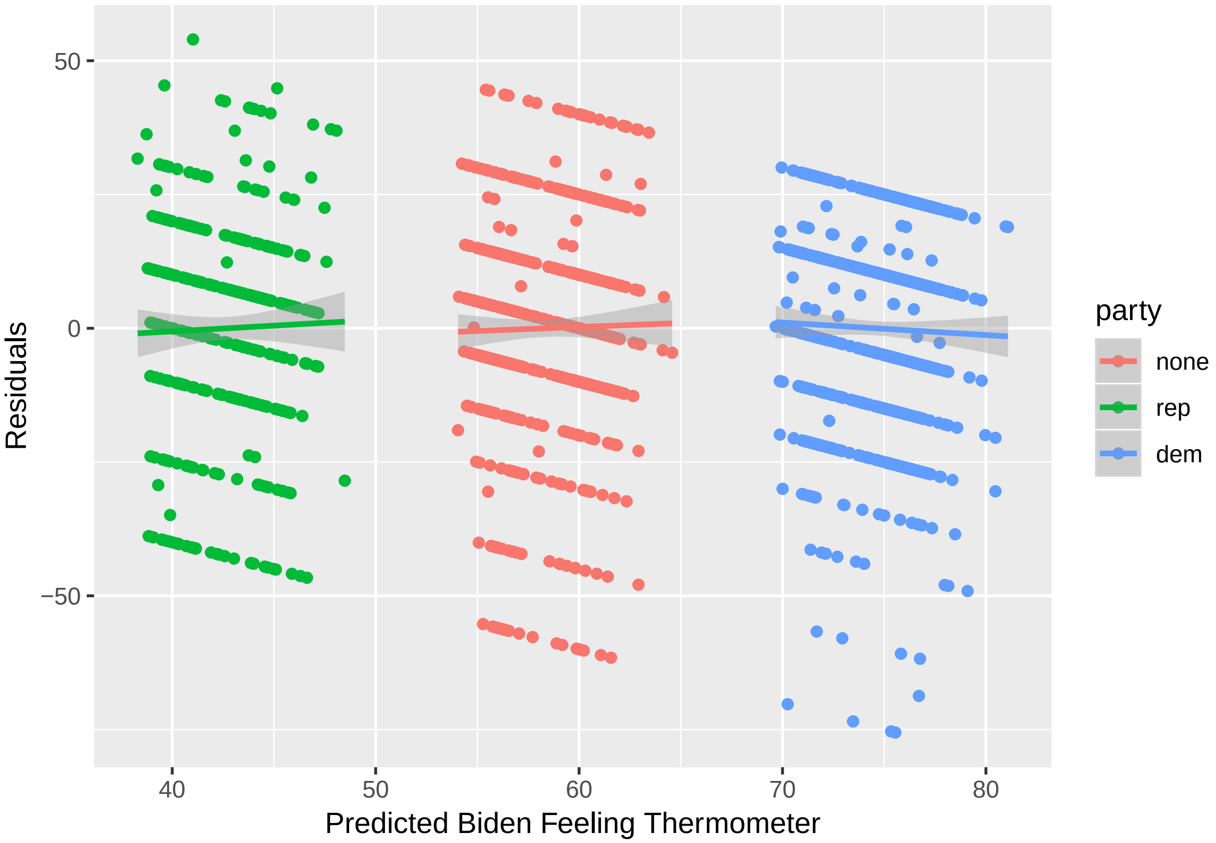
```
##
## Call:
## lm(formula = biden ~ age + female + educ + dem + rep, data = biden)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.546 -11.295   1.018  12.776  53.977
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.81126    3.12444  18.823  < 2e-16 ***
## age           0.04826    0.02825   1.708   0.0877 .
## female        4.10323    0.94823   4.327 1.59e-05 ***
## educ         -0.34533    0.19478  -1.773   0.0764 .
## dem          15.42426    1.06803  14.442  < 2e-16 ***
## rep         -15.84951    1.31136 -12.086  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.91 on 1801 degrees of freedom
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2795
## F-statistic: 141.1 on 5 and 1801 DF,  p-value: < 2.2e-16
```

a. The relationship has changed slightly.

b. This model explains 0.28% of the variance in the Biden feeling thermomter, which is better than the previous one.

c.

```
biden$fitted3 = fitted(model3)
biden$residual3 = residuals(model3)
biden$party = interaction(factor(biden$rep), factor(biden$dem))
levels(biden$party) = c('none', 'rep', 'dem', 'both')
biden$party = droplevels(biden$party, exclude = 'both')
ggplot(biden, aes(fitted3, residual3, color = party)) + geom_point(aes(color = party)) + stat_smooth(ae
labs(x = "Predicted Biden Feeling Thermometer",
     y = "Residuals")
```



The average of the predicted Biden feeling thermomoter among three parties are closer. However, none of the predicted values overlap acorss each other.

5.

```
biden2 = biden %>%
  filter(dem + rep != 0)

model4 = lm(biden ~ factor(female) * factor(dem), data = biden2)
summary(model4)
```

```
##
## Call:
## lm(formula = biden ~ factor(female) * factor(dem), data = biden2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.519 -13.070   4.223  11.930  55.618
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      39.382      1.455  27.060  < 2e-16 ***
## factor(female)1                   6.395      2.018   3.169  0.00157 **
## factor(dem)1                     33.688      1.835  18.360  < 2e-16 ***
## factor(female)1:factor(dem)1     -3.946      2.472  -1.597  0.11065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.42 on 1147 degrees of freedom
## Multiple R-squared:  0.3756, Adjusted R-squared:  0.374
## F-statistic:   230 on 3 and 1147 DF,  p-value: < 2.2e-16
```

```
pred_model4 = tibble(female = c(1, 1, 0, 0), dem = c(1, 0, 1, 0)) %>%
  bind_cols(as_tibble(predict(model4, . ,
                              interval = "confidence", level = 0.95)))
pred_model4
```

```
## # A tibble: 4 x 5
##   female   dem   fit   lwr   upr
##    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      1     1  75.5  73.8  77.3
## 2      1     0  45.8  43.0  48.5
## 3      0     1  73.1  70.9  75.3
## 4      0     0  39.4  36.5  42.2
```

From this model, we can learn that for both female and male, the Biden feeling thermometer rate for Democrats is higher than Republican. Nevertheless, for both Democrat and Republican, the female rate is slightly higher than male.